基于贝叶斯隐马尔可夫模型的疾病演变过程分析方法研究

# 摘要

慢性病通常随着时间的推移进展缓慢，对慢性病的演变过程进行分析可以帮助医生更好地了解和分析患者的疾病进展，实现慢性病相关症状的早期发现、早期预防、个性化诊疗和精准预后，是为慢性病提供更好的全程管理的基础和关键。传统的疾病演变分析是通过队列实验来进行的，但这些研究往往仅针对单一目标疾病及几个精心挑选的风险因子，基于严格的入组和出组条件，存在着患者特征项纳入不全的现象，且难以处理数据缺失的问题。

近年来，随着电子健康记录(Electronic Health Records, EHR)等医疗信息系统的快速发展，大量研究开始利用数据挖掘技术分析电子病历数据，进行疾病演变过程建模（Disease progression modeling, DPM），挖掘疾病的演变模式。相比传统的队列研究，DPM没有严格的入组出组条件，且EHR中记录了更丰富的患者就诊信息，便于纳入潜在的风险因子。然而，目前此类研究仍然存在以下不足：1）现有的模型往往只考虑了一种单一固定的疾病演变模式，而在真实世界中不同的患者可能有不同的演变模式和速度；2）现有模型没有充分利用EHR数据，而只是用到了其中小部分类型单一的患者特征。

针对以上研究的不足，本论文提出了一种基于贝叶斯隐马尔可夫模型（Bayesian Hidden Markov model, BHMM）的概率生成模型，对疾病演变过程进行建模。本论文提出的模型不局限于单一的疾病演变轨迹，允许疾病状态向任何其他状态转变。并且训练模型的数据包含了实验室检查、用药及手术等各类患者信息，充分利用了EHR数据。最后为了方便临床医生理解和使用模型，设计并实现了DPM展示系统。

将本论文提出的模型运用到从某三甲医院心内科收集到的2102份心力衰竭患者的EHR数据进行实验，实验结果表明：1）相比传统的DPM，本论文提出的模型能够更好地理解和代表患者的疾病演变轨迹。2）本论文提出的模型可以以较高的准确率来预测患者未来的临床表现。

**关键词**：疾病演变过程，贝叶斯隐马尔可夫模型，电子健康记录，心衰

# 目录

[摘要 I](#_Toc56543192)

[目录 II](#_Toc56543193)

[1 绪论 1](#_Toc56543194)

[1.1 引言 1](#_Toc56543195)

[1.2 研究现状 2](#_Toc56543196)

[1.2.1 队列研究 2](#_Toc56543197)

[1.2.2 基于EHR的研究 4](#_Toc56543198)

[1.3 研究目标与内容 6](#_Toc56543199)

[1.3.1 研究目标及任务 6](#_Toc56543200)

[1.3.2 论文内容安排 6](#_Toc56543201)

[2 数据统计分析 8](#_Toc56543202)

[2.1 基线数据统计 8](#_Toc56543203)

[2.2 案例分析 8](#_Toc56543204)

[2.3 本章小结 8](#_Toc56543205)

[3 基于贝叶斯隐马尔可夫模型的疾病演变过程建模 9](#_Toc56543206)

[3.1 基础知识 9](#_Toc56543207)

[3.1.1 贝叶斯网络 9](#_Toc56543208)

[3.1.2 隐马尔可夫模型 10](#_Toc56543209)

[3.2 模型构建 12](#_Toc56543210)

[3.2.1 疾病演变过程建模问题的定义 12](#_Toc56543211)

[3.2.2 基于贝叶斯隐马尔可夫模型的疾病演变过程建模 13](#_Toc56543212)

[3.3 实验和分析 17](#_Toc56543213)

[3.3.1 实验设置 18](#_Toc56543214)

[3.3.2 结果与分析 18](#_Toc56543215)

[3.4 模型评估 27](#_Toc56543216)

[3.4.1 评估方案设计 27](#_Toc56543217)

[3.4.2 结果与分析 27](#_Toc56543218)

[3.5 本章小结 28](#_Toc56543219)

[4 展示系统设计与实现 29](#_Toc56543220)

[4.1 整体框架 29](#_Toc56543221)

[4.2 功能设计 29](#_Toc56543222)

[4.3 系统实现 29](#_Toc56543223)

[5 总结与展望 29](#_Toc56543224)

[参考文献 30](#_Toc56543225)

# 绪论

## 引言

慢性病，也称为非传染性疾病，是指病情持续时间长、发展缓慢的疾病。慢性病的四个主要类型为：心血管疾病（如心脏病发作和中风）、癌症、慢性呼吸道疾病（如慢性阻塞性肺病和哮喘）以及糖尿病。根据世界卫生组织(WHO, World Health Organization)最新发布的《2020年世界卫生统计》显示，慢性病患病率近年来不断增加，引起的死亡总人数也不断增加，这种趋势遍及全球所有人群。在全球，有18%的人因心血管疾病、癌症、糖尿病、慢性呼吸系统疾病中的任一种而过早死亡[1]。2017年，全球共有5600万人死亡。70%以上的人死于非传染性慢性病。

心血管疾病(CVD, Cardiovascular Disease)是四大慢性病中危害最大，致死率最高的。心血管疾病是心脏和血管疾患引起的，包括冠心病（心脏病发作）、脑血管疾病（中风）、血压升高（高血压）、周围动脉血管疾病、风湿性心脏病、先天性心脏病和心力衰竭。根据WHO报道，心血管疾病是全球的头号死因：每年死于心血管疾病的人数多于任何其它死因。2016年，估计有1790万人死于心血管疾病，占全球死亡总数的31％。其中85%死于心脏病和中风。四分之三以上的心血管疾病死亡发生在低收入和中等收入国家。非传染性疾病导致的1700万七十岁以下死亡中，有82%发生在低收入和中等收入国家，37%由心血管疾病造成**错误!未找到引用源。**。

我国心血管病防治工作在取得初步成效的同时，也面临新的严峻挑战。根据《中国心血管病报告2019》[3]，目前，心血管病死亡占城乡居民总死亡原因的首位，农村为 45.91%，城市为 43.56%。如图1.1所示，从1990年到2017年心血管病死亡率呈总体上升趋势。其中从2006年开始，城乡心血管死亡率均急速上升，2017 年农村心血管病死亡率为 311.88/10 万，其中心脏病死亡率为城市

心血管病死亡率为 268.19/10 万。中国心血管病负担日渐加重，已成为重大的公共卫生问题，对人民的健康造成了巨大的威胁，防治心血管病刻不容缓。

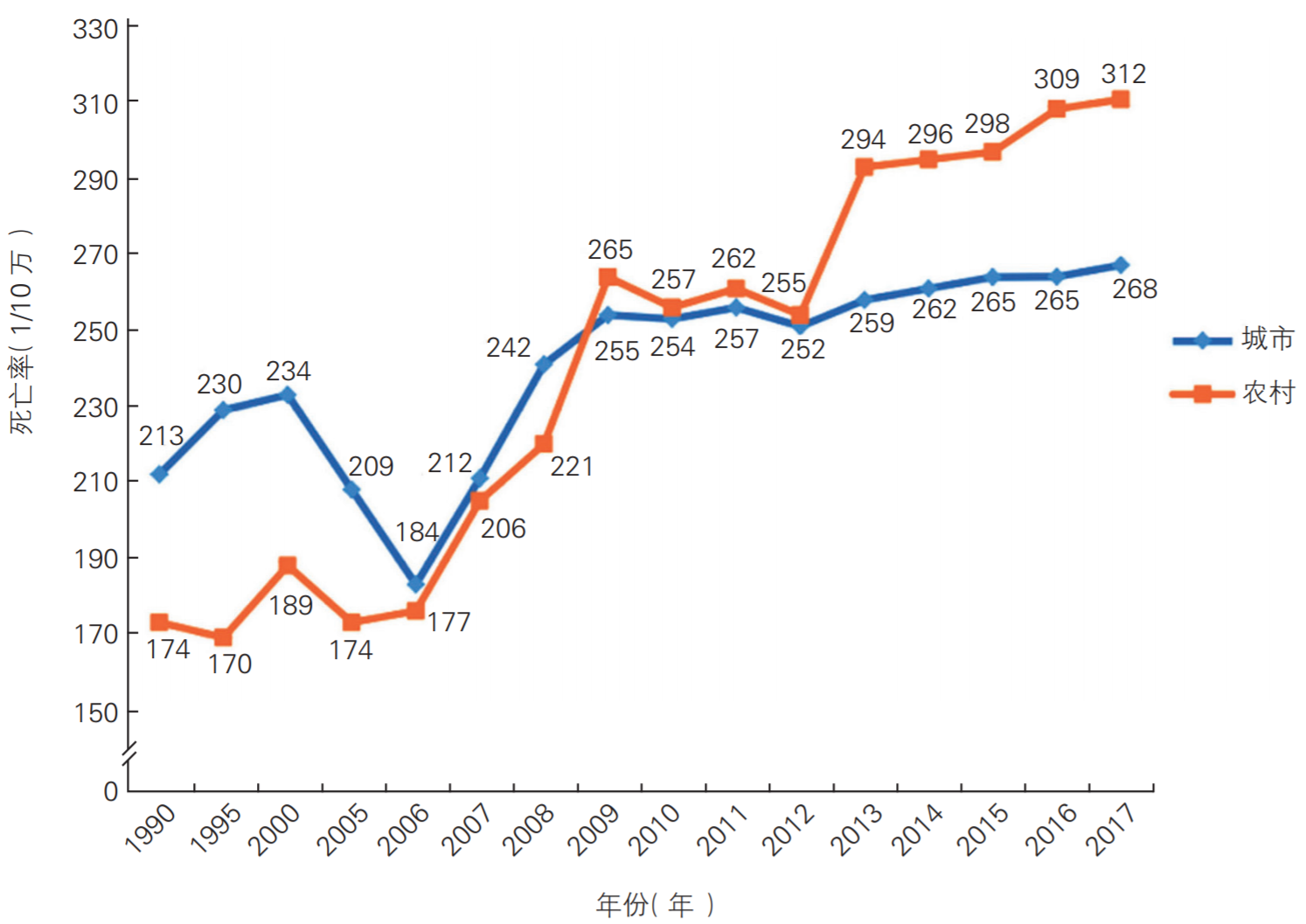


图1.1 1990-2017年中国城乡居民心血管病死亡率变化

疾病演变过程分析（Disease Progress Analysis）即对心血管等慢性病的发展过程进行分析，从而深入了解整个患者治疗轨迹，帮助医生在早期阶段跟踪慢性疾病的发展，有助于早期发现和预后,从而促进有效的医疗服务和更好的疾病管理[4]。同时，随着各大医院信息化建设进程的不断推进，医院的信息系统存储了大量的电子健康记录(EHR, Electronic Health Record)数据，包括大量的患者检查、检验等临床诊疗数据。为疾病演变过程分析提供了丰富的数据来源。因此，可以利用数据挖掘技术分析电子病历数据，挖掘疾病的演变模式，为临床医生评估疾病的进展，进而开展针对性的医疗干预[5]。

## 研究现状

疾病演变过程分析方法主要有队列研究和基于电子病历数据两种。下面将分别描述：

### 队列研究

在临床医学研究中，慢性病演变分析研究一般建立在临床队列实验的基础上。队列实验指的是将某一特定人群按是否暴露于某可疑因素或暴露程度不同分为不同的亚组，追踪观察两组或多组成员结局事件（如疾病）发生的情况，比较各组之间结局发生率的差异，从而判定这些因素与该结局之间有无因果关联及关联程度[11]。对于针对疾病演变过程分析的队列实验，首先会制定入组和出组标准来纳入研究对象，然后通过长期随访患者群，得到若干患者特征的长期纵向数据，利用多元逻辑回归分析等方法，筛选出几个独立的临床特征（风险因素），以其分值之和将其分为不同危险等级的患者，进而描绘出患者的疾病演变轨迹，并做出预后评价，指导治疗。针对其中的入组与出组标准、患者特征及结局事件的论述如下[12]：

* 入组与出组标准：入组标准即对患者的症状、诊断、合并症及超声心动图等检查检验结果做出具体的要求，满足要求的患者纳入队列；出组标准主要是对会干扰实验结果准确性的因素进一步的排除，比如患者的既往病史及入院后所接受的治疗，符合出组标准的患者会被排除。通过入组和出组标准，队列研究可以严格控制研究对象的特征和质量，保证实验结果的有效性。
* 患者特征：一组临床指标，用来量化描述入组对象。通常在初期的实验设计阶段，根据研究目的，确定需要长期随访来收集的患者数据项，如人口特征、体征、症状、用药以及检查检验相关结果。
* 结局事件：即对入组对象随访结束的标志。如死亡、再发脑梗/心梗等。

已经有大量针对疾病演变过程分析的队列研究。Richard等人[6]根据美国、欧洲和南美共四个临床中心的随机对照临床实验，比较了非清髓性造血干细胞移植与疾病改良疗法对疾病进展的影响。Chen等人[7]提出了一个分类模型用来模拟心律失常性心肌病（AC），以发现一组不同的潜在发病机制，采用的队列包含60名具有不同遗传背景的AC患者。Annul等人[8]开展了一项纵向观察研究，以确定前房角宽度是否在窄角法向性青光眼的进展中起到相关作用。Eduardo等人[9]研究了帕金森病中自主神经功能障碍发作时间与疾病进展和生存的关系。Ito等人[10]提出了一个基于文献元分析的模型来描述轻中度阿尔茨海默病患者的纵向变化。

然而这些研究都具有以下所述的不足：

* 严格的入组和出组标准使得研究结果不具有普适性：几乎所有沿这方面的研究都倾向于针对某一种特定的疾病，研究目标也非常具体，如不同干预方法的治疗效果。在入组和出组条件的严格限制下，很多年龄较大或较小、同时患有多种合并症的患者不能进入队列，而在真实的临床环境中，这些情况是非常常见的。因此队列研究收集到的数据与真实世界存在一定的差异[12]。
* 患者特征项纳入不全：这些队列研究采用的是精选挑选的少量患者特征，无法纳入新发现或潜在的疾病特征。随着医学技术的发展，越来越多的潜在疾病特征不断被发现和重视，如NTproBNP、eGFR等指标可能对心衰的发生和发展过程有一定的判别价值，但并未包含在NYHA心功能分级等常见的心衰疾病演变分析模型中。
* 无法处理数据缺失的问题：此类队列研究依赖完备的长期纵向数据集来建模，无法处理含有缺失值的患者。然而，在真实的临床环境中，患者记录中经常含有缺失值，因此限制了此类队列研究模型的使用[14]。

### 基于EHR的研究

随着信息技术在医疗领域内的广泛应用，大量的患者临床诊疗数据被记录并保存在EHR中。EHR在患者治疗过程中记录了各种类型的患者信息（例如症状，生命体征，实验室检查结果等）和治疗干预措施（例如药物，手术，检查等），因此隐藏了尚未被发现的大量关于疾病演变过程的信息[15]。从这个意义上讲，EHR具有独特的潜力，可用于建立疾病演变过程的数据挖掘或统计模型，然后通过这些模型帮助医生了解疾病进展并提供有效的健康服务，改善单个患者的治疗实践。相对于队列研究，由EHR驱动的模型没有严格的纳入和排除标准，使用的数据更符合临床的真实情况，也能够方便地纳入新的患者特征和排除具有缺失值的记录[16]-[18]。目前已经有一些由EHR驱动的疾病演变过程建模（DPM，Disease Progression Modeling）方法，例如：

Cohen等人[19]对在重症监护病房每分钟收集到的45个生理、临床和治疗变量进行分层聚类，以确定患者状态，识别重症监护病房中的基本疾病进展模式。结果表明，层次聚类技术可用于可视化复杂的多变量数据，并可为重症患者的护理提供新的见解。

Zhou等人[20]-[21]提出了一个多任务学习框架，通过临床评分（ADAS-Cog和MMSE）来预测疾病的进展。对来自阿尔茨海默症神经影像学倡议（ADNI）数据集进行广泛的实验研究，表明提出模型的有效性。

Sukkar等人[25]提出了基于隐马尔可夫模型（HMM, Hidden Markov Model）框架的疾病进展模型。针对阿尔茨海默症的生物标记物，使用ADNI数据集以无监督的方式训练了HMM，结果表明，受过训练的HMM能够比当前定义的临床阶段更精细地模拟疾病进展。

Jackson等人[24]建立了一个多阶段隐马尔可夫模型，用于在连续时间内错误分类的马尔可夫过程。可同时估算疾病各个阶段之间的转移概率和阶段被错误分类的概率，并将其应用于动脉瘤筛查研究，根据主动脉直径的连续范围按严重程度对疾病进行分级。

然而，存在一些不足限制了这些模型的通用性。首先，尽管疾病的演变是一个连续的过程，但是观察数据只在一系列离散的时间点收集到，并且这些时间点之间的间隔也是不相等的；其次，单个患者的记录往往是不完整的，仅仅只涵盖了整个疾病演变过程的一小部分，必须将大量患者的记录“拼接”在一起来推出完整的演变过程。针对这些挑战，已经有部分研究提出了新的解决方法：

Wang等人[22]提出了一种连续时间隐马尔可夫模型（CTHMM），通过合并症（如肾病、糖尿病）的发作来推断慢性阻塞性肺疾病（COPD）的进展，可以从不完整、不规则、离散的病人记录中学习完整的、连续的疾病演变轨迹。特别是，他们的工作将患者记录建模为时间序列，该时间序列可以在进展路径的任何点开始和结束。

Sun等人[23]使用了CTHMM和主题模型分析技术，对亨廷顿（HD，Huntington’s Disease）病人的发展阶段进行了划定，并且对每个阶段的主要特征进行了分析。实验结果表明他们的模型相比临床采用的分期标准涵盖了更精细的阶段划分，并且能够定量描述临床诊断前后的复杂变化；

但是，这类研究依然存在两点不足：1）仅侧重于发现目标疾病的单一演变轨迹，通过对马尔可夫过程加以约束使其仅能向前转换，即疾病状态只能从早期病情轻微的阶段转移到后期病情恶化的阶段，禁止从后期转移到早期。然而，对于许多疾病来说，大多数患者不太可能遵循单一的疾病演变轨迹，而是有多种情况。2）这些研究没有充分利用电子病历数据，而只是利用了其中的一小部分，如合并症或者根据现有评分模型选取的少量风险因子，鲜有研究将各种类型的数据综合起来构建更为全面、准确的模型。

## 研究目标与内容

### 研究目标及任务

本论文针对以上基于EHR的疾病演变过程分析的两点挑战，提出了一种全新的贝叶斯隐马尔可夫模型（BHMM，Bayesian Hidden Markov Model）。该模型不限制疾病状态只能从早期转移到晚期，而允许跳回到之前的状态，即本论文提出的模型可以根据真实数据所隐含的信息，学习到各种可能的疾病状态转移路线。

此外，为了充分利用EHR数据，该模型的输入包含了EHR中的各种患者信息，如体征、用药、手术及实验室检查等各类信息，使建模结果更具有说服力。

本论文的研究内容主要包括以下几个方面：

* 数据统计与分析。TODO
* 基于贝叶斯隐马尔可夫模型的疾病演变过程建模（BHMM-DPM）。本论文拓展了最新的BHMM，将其应用到DPM领域。并使用心衰患者的EHR数据进行实验验证其可行性和有效性。
* 疾病演变过程分析系统设计与实现。为了便于临床医生理解和使用，本论文以心衰患者为例，设计并实现了心衰演变过程分析系统，在线查看心衰的各个隐藏疾病状态及推测患者的状态转移路线，为医生制定治疗措施和评估干预效果提供参考意见。

### 论文内容安排

本论文一共分为五章。

第一章——绪论。介绍了疾病演变过程分析及建模的概念及研究意义。讨论了基于队列研究和基于EHR的疾病演变过程分析研究的现状，分析其不足之处，并据此说明本论文主要解决的问题及研究内容。

第二章——数据统计与分析。对本论文实验所用的数据进行全面的统计分析。TODO

第三章——基于贝叶斯隐马尔可夫模型的疾病演变过程建模。首先介绍了贝叶斯理论和隐马尔可夫模型的原理，然后介绍本论文提出的BHMM-DP模型的定义。使用心衰患者EHR数据集对模型进行训练，将其结果和朴素贝叶斯模型进行对比验证BHMM-DP的有效性。

第四章——疾病演变过程分析系统设计与实现。基于本论文提出的模型，设计并实现了心衰演变过程分析系统，对心衰患者所处的阶段进行预测，帮助医生制定合理的治疗计划。

第五章——总结与展望。总结了本论文所做的工作和取得的成果，讨论了现阶段工作的不足之处，并对后续研究工作进行了展望。

# 数据统计分析

本论文选用心力衰竭（以下简称心衰）作为典型的心血管病的代表来分析其演变过程，采用的数据集是来自国内某三甲医院心内科的2102份心衰患者的HER数据，包含了患者的体征信息、检查检验信息、用药情况等各类数据。为了充分利用HER数据，本论文采用了七大类型的综合数据作为患者特征。本章将对这些数据进行统计分析，并通过几个患者样本多次入院组成的心衰演变轨迹来

分析心衰演变过程的特点。

## 基线数据统计

本论文采用的EHR数据集包含2102份心衰患者的13546次入院记录，记录的患者信息包括基本信息、实验室检查结果、用药信息、手术信息、超声心动图检查结果、心功能分级及患有其他合并症的情况这七类。本节将分别对这七类数据进行简要的统计分析，以展示各类数据在患者中的分布情况。需要说明的是，

对于一些有多次检查结果的患者特征（如各项体征和实验室检查结果），本节选取第一次的结果作为患者的基线数据，代表患者在就诊初期时的状态。

1. 基本信息

表2.1 基本信息



为了全面深入地研究疾病演变的过程，在数据预处理阶段筛选掉了入院次数小于3次的患者，最终数据集中所有患者的平均入院次数为6.44次。患者平均年龄为62.25岁，说明老年患者居多，但样本集涵盖了各个年龄段的患者，因此标准差较大。至于血压、BMI等指标，平均值处于正常偏高的水平。

## 案例分析

## 本章小结

# 基于贝叶斯隐马尔可夫模型的疾病演变过程建模

本章提出了一种新颖的概率生成模型来对患者的疾病演变过程进行建模，该模型将患者的疾病状态视为马尔可夫过程中一个隐藏的变量，通过已知的患者诊疗数据反推出疾病状态。并且使用第二章中的EHR数据进行了模型的训练和相关实验，并对模型的有效性进行评估。

## 基础知识

在详细介绍本论文提出的模型BHMM-DP之前，先介绍一下该模型应用到的基础知识：贝叶斯网络和隐马尔可夫模型。

### 贝叶斯网络

贝叶斯网络(Bayesian Network) 于1985年由Judea Pearl[26]首先提出，又称信念网络(Belief Network)或有向无环图(DAG, Directed Acyclic Graphical)模型，是一种概率图模型。它是一种模拟人类推理过程中因果关系的不确定性处理模型，网络拓扑结构是一个有向无环图，表示一组随机变量及其相互之间的条件依赖。

贝叶斯网络使用有向无环图中的节点代表随机变量，它们是可以观察到的变量，或潜在变量、未知参数等。连接两个节点的有向边代表这两个随机变量具有因果关系（或非条件独立），其中有向边起点处的节点称为父节点，而末尾的节点称为子节点；而两个节点之间若没有有向边相连就称这两个随机变量彼此间为条件独立。例如，假设节点E直接影响到节点H，即E→H，则用从E指向H的箭头建立节点E到节点H的有向边(E,H)，权值(即连接强度)用条件概率P(H|E)来表示，如下图所示：



图3.1 贝叶斯网络中具有因果关系的两个节点

贝叶斯网络的数学定义如下：

令 G= (I,E) 表示一个有向无环图，其中 I 代表图中所有的节点的集合，而 E 代表所有有向边的集合，令X= (Xi)i∈I 为其有向无环图中的某一节点 i 所代表的随机变量，若节点 X 的联合概率分布可以表示成：

(3.1)

则称 X 为相对于有向无环图 G 的贝叶斯网络，其中 pa(i)表示节点 i 之「因」, 或称pa(i)是i的parents（父母）。

对任意的随机变量，其联合分布可由各自的局部条件概率分布相乘而得出：

一个简单的贝叶斯网络如下图所示：

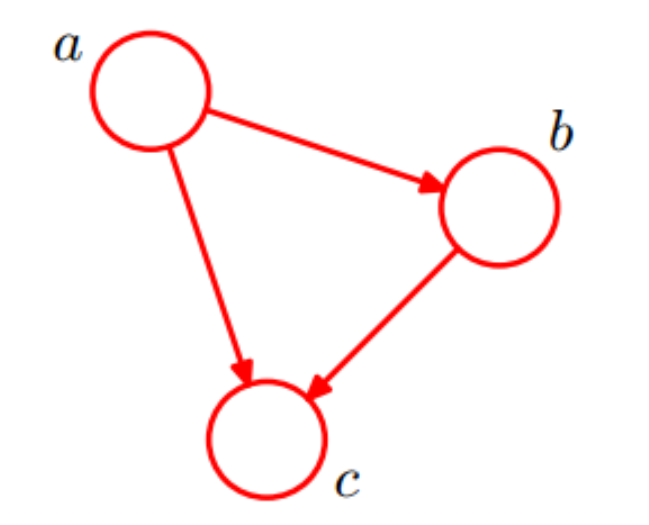


图3.2 一个简单的贝叶斯网络

因为a影响b，a和b影响c，所以a,b,c的联合分布概率为：

(3.3)

### 隐马尔可夫模型

隐马尔可夫模型（Hidden Markov model, HMM）是一种关于时序的概率模型，描述由一个隐藏的马尔可夫链随机生成不可观测的状态序列(state sequence)，再由各个状态生成一个观测而产生观测序列（observation sequence）的过程。HMM由Leonard E. Baum等人[27-29]于20世纪提出，作为一种经典的机器学习模型，它在语音识别、故障诊断及模式识别等领域得到了广泛的应用。

首先介绍一下马尔可夫链。马尔可夫链也叫马尔可夫过程，是在贝叶斯网络中概率图退化成线性链时的特殊情况，如下图所示：



图3.3 马尔可夫链

马尔可夫链描述了状态空间中从一个状态到另一个状态的转换的随机过程。这个过程需要满足马尔可夫性质：当前状态只和上一个状态有关，而和之前的任何其他的状态无关。即在图3.3中，的分布状态只和有关，和其他变量条件独立。

HMM是在马尔可夫链的基础上发展而来的，可以看作一个双重的随机过程。在基本的马尔可夫随机过程中，状态序列就是可以直接观察到的观测序列。而HMM的状态序列是不可直接观测到的，观测值是由“隐藏的”状态随机产生的[30]。HMM的结构如图3.4所示：

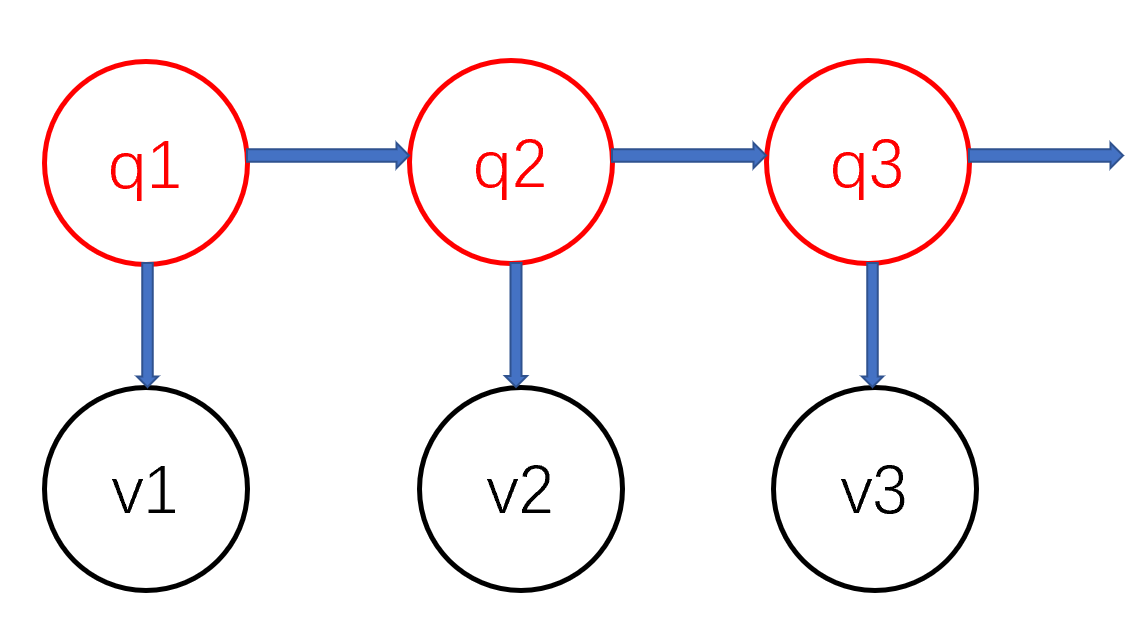


图3.4 HMM结构示意图

设Q = {,…,}是所有可能的状态的集合，V = {,…,}是所有可能的观测的集合，其中N是所有可能的状态数，M是所有可能的观测数。设I = {,…,}是长度为T的状态序列，O = {,…,}是对应的观测序列。HMM由初始状态概率分布**π**、状态转移概率分布A，观测概率分布B这三个参数确定，记为λ= (π, A, B)，下面分别对每个参数进行说明：

* 初始状态概率分布**π={}，**其中

(3.4)

代表时刻t=1时处于状态的概率。

* 状态转移矩阵A={}，其中

表示在时刻t处于状态的条件下，在时刻t+1转移到的概率。

* 观测概率矩阵B={}，其中

表示在时刻t处于状态的条件下生成观测的概率。

HMM的三个基本问题为：

1. 概率计算问题：给定模型λ= (π, A, B)和观测序列O = {,…,}，计算在模型λ下出现该观测序列的概率P(O|λ)，是最基本的问题。
2. 学习问题：已知观测序列O = {,…,}，估计HMM的参数λ= (π, A, B)，使得在该模型参数下观测序列O出现的概率P(O|λ)最大。这个问题也是本论文提出的模型要解决的问题，实际上是通过已知的观测序列训练出一个HMM。
3. 预测问题：也称为解码问题，已知HMM参数λ= (π, A, B)和观测序列O = {,…,}，求一个状态序列I = {,…,}使得条件概率P(I|O)最大。即产生给定观测序列对应的最有可能的状态序列。

## 模型构建

本节会详细地介绍模型的细节，分为两部分：

* 临床观察项和疾病演变过程的符号表示
* 模型的图形表示和算法步骤

### 疾病演变过程建模问题的定义

在本论文中，用集合来表示|D|个患者的EHR记录，用集合代表在中出现的所有临床观察项，包括患者的人口统计学、体征、实验室检查结果、用药及手术等信息。每个EHR d()对应一个特定的患者，由一组患者入院记录组成，即，其中包含一系列临床观察项，它们描述了患者在第t次入院就诊时的疾病状态，即，其中，是患者d在整个治疗轨迹中入院的总次数。

需要特别指出的是，多次入院之间存在着某种顺序依赖性。具体来说患者每次入院的EHR，描述了患者在该次入院时所处的疾病状态，这是一个隐含的抽象变量，可以表示为关于一组临床观察项的多项分布。实际上，疾病状态z就是上一节介绍的隐马尔可夫模型（HMM）中的“隐藏”的变量。也就是说，将患者的多次入院看作一个隐马尔可夫过程，每次入院对应一个隐藏的状态，而每次入院的EHR中记录的一组临床观察项则是该隐藏状态对应的可观测项。

利用一组EHR集合中某些临床观察项经常同时出现的信息，无监督的学习方法可以自动发现代表某个疾病状态的一组高度相关的临床观察项。而一旦一组高度相关的临床观察项从EHR中被提取出来，医学研究者就可以给它们分配有意义的标签用于进一步的应用中，例如本论文中的疾病演变过程分析。本论文要解决的问题定义如下：

**问题** 给定一个EHR集合，其中每个EHR 包含一系列入院记录，每次入院 具体表示为在该次入院期间产生的一组临床观察项，包括患者的症状、体征即用药等干预措施，这些信息反映了病人在该次入院所处的隐藏的疾病状态。假设所有的患者样本共享K个隐藏疾病状态。则疾病演变过程建模（DPM）的目的就是从每个患者治疗轨迹的EHR中识别出每次入院所处的疾病状态以及多次入院之间疾病状态转移的路线。

### 基于贝叶斯隐马尔可夫模型的疾病演变过程建模

针对上一小节中提出的问题，本节提出了一种基于贝叶斯隐马尔可夫模型的概率生成模型，记为BHMM-DP，其概率图如图3.5所示，相关变量的符号如表3.1所示：

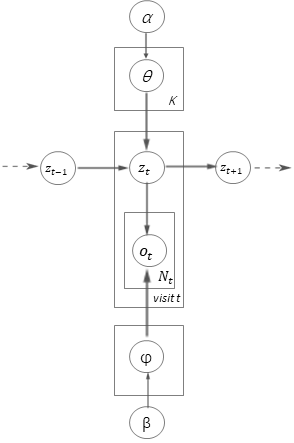


图3.5 基于贝叶斯隐马尔可夫模型的疾病演变过程模型（BHMM-DP）

表3.1 本研究中使用的符号

|  |  |
| --- | --- |
|  | 狄利克雷分布（Dirichlet）的先验参数 |
| *K* | 疾病状态的数目 |
| *O* | 临床观察项的数目 |
|  | 一个病人的治疗轨迹 |
|  | 患者的第*t*次入院 |
|  | 第*t*次入院期间产生的临床观察项 |
|  | 疾病状态转移的分布概率 |
|  | 第*t*次入院对应的疾病状态 |
|  | 特定于临床状态z的临床表现的多项式分布 |

如图3.5所示，假设在一个患者的治疗轨迹中，第t次入院对应的疾病状态由该次入院产生的临床观察项和上一次入院（即第t-1次）所处的疾病状态决定。这个假设遵循一阶马尔可夫性质，因此可以应用一阶马尔可夫链结构来出来DPM问题。

具体来说，给定一个患者样本，包含次入院记录，其疾病演变过程可以表示为：

(3.7)

(3.8)

(3.9)

(3.10)

其中公式（3.7）表示疾病状态服从多项分布，指的是当上一次入院所处的疾病状态为时，第t次入院的疾病状态转移概率分布。在公式（3.8）中，指的是在d(t)中记录的第j个临床观察项，指的是当疾病状态为时临床观察项的概率分布。此外，、分别服从参数为α、β的Dirichlet分布，且本研究采用对称的先验参数。

图3.5所示BHMM-DP的生成过程为：

给定一个患者的治疗轨迹d和其中的第t次入院，参数为α的狄利克雷分布会产生一个疾病状态转移分布的先验分布（即θ），参数为β的狄利克雷分布会产生一个临床观察项的先验分布（即）。因此，本次入院对应的疾病状态由疾病状态转移概率分布和上一次入院对应的疾病状态所确定，最终一个临床观察项o由本次入院对应的疾病状态下临床观察项的多项分布产生。因此，BHMM-DP模型中观测到的以及隐藏的变量对应的联合分布概率如下所示：

. (3.11)

为了评估出BHMM-DP的模型参数，最优方法是使用最大似然估计，使得给定的临床观测项出现的概率最大。然而，很难直接通过数学方法来最大化似然函数，解决方法是使用近似推断，本论文中采用了吉布斯采样（Gibbs Sampling）方法【文献】。首先，从公式（3.11）中可以推导出每个患者样本每次入院d(t)的全条件分布（full conditional distribution），如下所示：

(3.12)

其中表示从中排除当前的入院记录d(t)，代表除了当前入院d(t)之外的所有入院对应的疾病状态。

接下来采用一个包含两个步骤的吉布斯采样方法来估计式（3.12）中的每个乘积项。第一步，是根据以下公式来采样疾病状态z：

(3.13)

(3.14)

式（3.13）中是指除了当前入院外，所有其他入院诊断中疾病状态为z而其上一次入院的疾病状态为的总次数，分母中的。式（3.14）同理，其中是一个指示函数，即如果为真则等于1，如果为假则等于零。

第二步则是根据以下后验概率得到在d(t)中的临床观察项：

(3.15)

其中满足：

其中指除了本次入院外在其他所有疾病状态为z的入院中出现临床观察项的次数。

当采样进行足够多的迭代次数之后，可以通过以下（3.17）和（3.18）所示的后验估计来估计参数和 ：

|  |
| --- |
| (3.17) |
| (3.18) |

算法1显示了训练BHMM-DP的伪代码：

|  |  |
| --- | --- |
| 算法1: 基于贝叶斯隐马尔可夫模型的疾病演变过程建模 | |
| **Input**: A EHR dataset , hyper-parameters , the number of disease states , and the number of iterations | |
| **Initialization step:** | |
|  | Initialize , and |
|  | For each patient sample |
|  | For each visit |
|  | Sample a disease state assignment |
|  | Increment = |
|  | For each clinical manifest |
|  | Increment |
|  | End For |
|  | End For |
|  | End For |
| **Gibbs sampling step:** | |
|  | For each iteration |
|  | For each patient sample |
|  | For each visit |
|  | Decrement = |
|  | For each clinical observational item |
|  | Decrement |
|  | End For |
|  | Sample a disease state assignment using Equation (3.12) |
|  | For each clinical manifest |
|  | Increment |
|  | End For |
|  | Increment = |
|  | End For |
|  | End For |
|  | End For |
| **Output**: Estimate the model parameters with sampling results as follows: | |
|  |  |
|  |  |

模型训练完后，一个重要的任务就是当给定患者样本d的一个新的入院记录d(t)时，根据训练好的模型推断除该次入院对应的疾病状态。具体而言，就是计算每一个疾病状态对应的概率，然后选择概率最大的作为该次入院对应的疾病状态。计算疾病状态的公式如下：

(3.19)

特别地，如果d(t)是第一次就诊，即没有上一次就诊d(t-1)，则认为。

## 实验和分析

本节将会使用第二章中的心衰数据集对上面提出的BHMM-DP进行训练，评估出模型参数，然后据此进行进一步的实验以深入分析心衰演变过程。实验包括：

* 疾病状态数目的选取
* 疾病状态的发现
* 疾病状态与合并症的关联性
* 疾病状态转移路线

### 实验设置

对于BHMM-DP中的超参数α和β，训练模型采用常用值α= 50/K，β= 0.01。完成实验的平台是Windows 10操作系统，处理器的型号为Intel® Core™ i5-8265U CPU @ 1.60GHz，运行内存为8GB。模型的训练和所有的实验都是采用Python语言实现。

### 结果与分析

**3.3.2.1疾病状态数目的选取**

**在上一节训练模型的算法1伪代码中，需要输入疾病状态的数目***K***。本论文选取使用最为广泛的方法——基于困惑度（perplexity）的方法来确定***K***的值。具体到本研究中，对于由*D*个患者的EHR数据集，困惑度定义为：**

**其中**指的是在患者d的疾病演变轨迹中出现过的临床观察项，分子中的可以表示为：

其中是患者d出现的临床观察项的总数目。一般而言，模型困惑度会随着*K*的增大而减小。而另一方面，如果*K*过大则会出现过拟合的现象。因此需要在模型的简洁性和拟合度之间做一个平衡[31]。

**本论文分别设置***K*={1, 3, 5, 7, 10, 13, 15, 20, 25, 30}来训练所提出的BHMM-DP，结果如图3.7所示，由图可知，刚开始时模型的困惑度随着疾病状态数量的增加迅速下降，但当*K*大于10时，困惑度的下降速度显著降低。因此，最终选取*K*=10。

**图3.6 困惑度随疾病状态数目的变化**

**3.3.2.2 疾病状态的发现**

如前所述，本研究提出BHMM-DP的目的就是利用EHR数据发现隐含在患者疾病演变轨迹中的疾病状态。为此，实验过程中计算了模型中的参数，然后列出了与每种隐藏的疾病状态高度相关的前40项临床观察项，并按降序排列的方式列在表3.2中，注意表中与实验室检查相关的临床观察项的后缀H、N、L分别代表该项检查指标偏高、正常和偏低，此外加粗显示的为不正常的观察项。

从表3.2可以看出，不同的心衰状态下对应的各个临床观察项出现的概率大不相同。在状态2下，异常指标主要是“脑利钠肽前体偏高”、“血浆D-二聚体测定偏高”及血细胞减少等，预示着心衰患者的病情处于中等的状态。状态6下绝大部分临床观察项都正常，“心源性入院”的概率偏高，意味着处于状态6下的心衰患者病情较轻。与此相反，状态8则对应最多的异常指标，除了“脑利钠肽前体偏高”外，还有“肾功能不全”、“高血压”和“慢性病”等合并症以及“利尿剂”等干预措施。所以心衰等慢性病病情复杂多变，患者的疾病演变轨迹应该是各种可能的疾病状态的组合，而不是某一种单一的演变模式。

图3.7显示了各个心衰状态出现在不同的患者中的次数。可见，出现这10个状态的人数差别不大，即每一种状态对应的临床观察项的组合在心衰患者的病情演变轨迹中都很常见。

表3.2 心衰状态-临床观察项表









图3.7 分配给每种疾病状态的患者数目

**3.3.2.3 不同疾病状态与合并症的相关性**

**本小节设计实验来探究合并症在不同疾病状态下的发生情况。具体来说，本实验选取了三种常见的合并症：高血压、冠心病及肾功能不全，分别统计了在所有就诊记录中各个疾病状态下每种合并症出现的次数，结果如图6.8-6.10所示。**

**不同状态下对应的合并症出现频率有明显区别，且不同的合并症与疾病状态之间的关系也互不相同。对于高血压，状态3和状态4对应的高血压出现次数最多，而状态2对应的高血压出现次数最少；而对于冠心病，则是状态3和状态6出现的次数最多，同样也是状态2出现的次数最少；至于肾功能不全，在状态4、**

**图3.8 高血压出现次数与疾病状态的关系**

**图3.9 冠心病出现次数与疾病状态的关系**

**图3.10 肾功能不全出现次数与疾病状态的关系**

**8和10中出现的次数远远高于其他状态。这与表3.2展示的各状态对应的临床观察项的概率基本一致，也反映出BHMM-DP模型的准确性。**

**3.3.2.4 疾病状态转移路线**

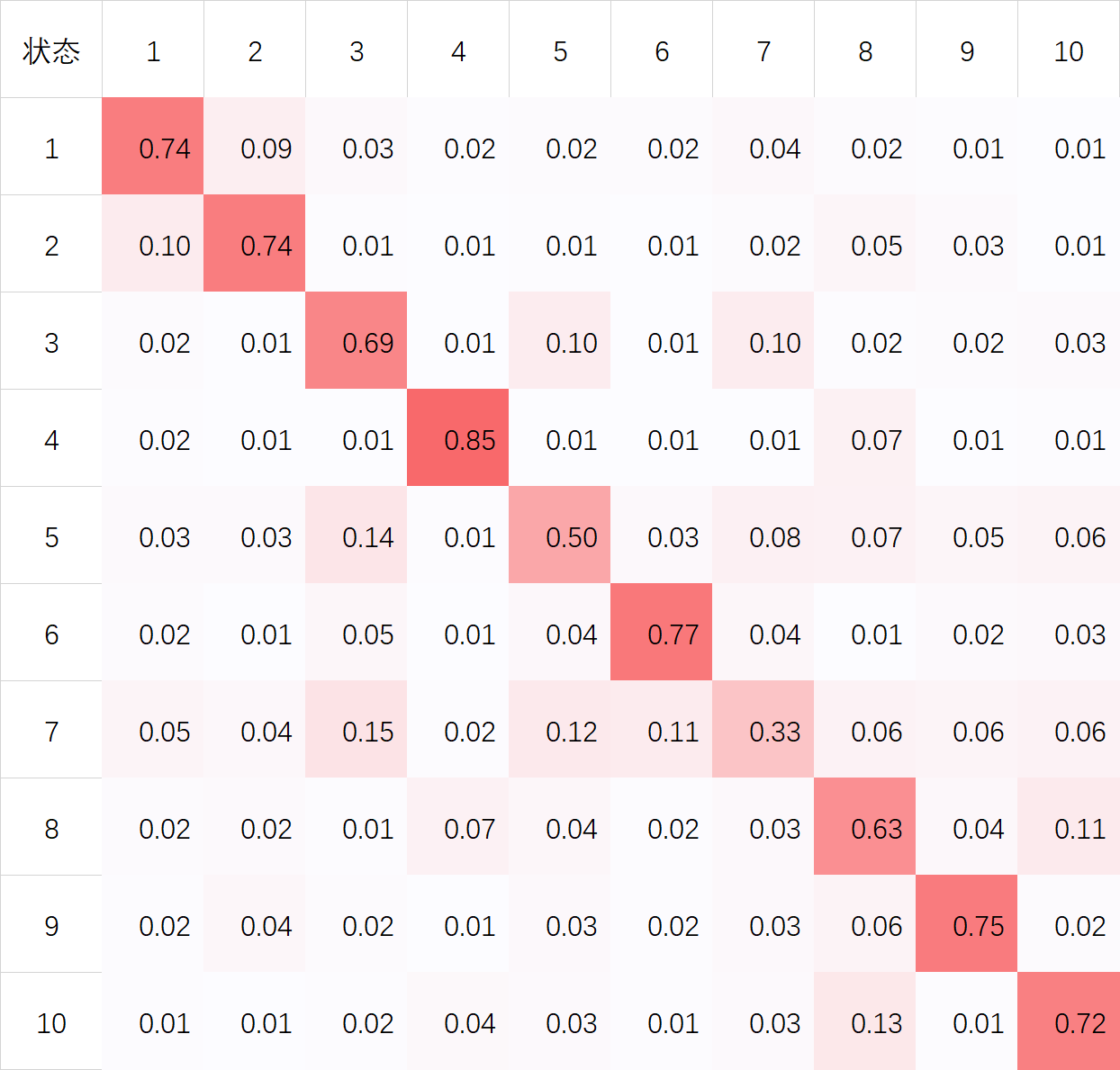
**图3.11 不同状态的患者数量随时间的变化**

**图6.11展示了不同状态的患者数量随时间的分布，z轴表示在特定就诊次数处于特定状态的人数。在最开始的几次入院就诊，大部分患者处于状态3、6或9，随后会转移到状态8和状态10。根据表3.2，与状态3、6和9相关的临床观察项大部分都比较正常，符合心衰早期症状。而状态8和10则包含了较多异常临床观察项，尤其是状态8，死亡概率最高。说明随着时间的推移，患者病情逐渐加重，部分患者经过治疗维持在某个状态，而部分患者病情持续加重，直至死亡。**

**此外，模型训练完后也得到了各个心衰状态之间的转移概率矩阵，如图6.12所示，对角线上的概率通常大于非对角线，这表明大多数患者在多次就诊之间会保持相同的状态，只有一小部分患者会转移到其他的状态。**

**为了直观地说明如何将本论文提出的模型运用于推断单个患者的疾病演变轨迹，本小节从实验数据集中选取了五个代表性的患者样本来分析他们的心衰演变过程。具体来说，根据3.2.2节中公式（3.19）来计算出每个患者每次入院最有**

**可能的状态，结果如图3.13所示。**

 ****

**图3.12 状态转移矩阵**

**图3.13 心衰患者个体状态转移路线**

**患者1第一次入院时处于状态1，根据表3.2，状态1各项临床观察项基本正常，是病情最轻微的状态，后续病情变重，由于心脏病发作第二次入院就诊，随后就转移到了状态3，并且接下来几次入院都保持状态3，说明病情被控制得较为稳定。与患者1相反，患者2一开始就患有心血管疾病（心力衰竭、心律失常），在三次入院后恶化为最严重的状态8，预后不乐观。这两名患者的病情演变分析反映了本论文提出的BHMM-DP模型在临床决策支持系统中的潜力，即为医生提供合理的预测和干预建议。**

## 模型评估

### 评估方案设计

为了定量评估提出模型的准确性，本小节提出了以下的评估方案和指标：

给定患者样本，取前次的就诊记录来拟合模型，即根据训练好的BHMM-DP模型参数和这次的EHR数据推断出第次入院对应的状态，然后用以下公式估计第次入院对应的临床观察项：

具体来说，首先计算出每个临床观察项出现的概率大小，然后根据概率进行排序，取前N个临床观察项，认为它们会在第次入院时出现。以真正的第次入院时的EHR数据作为金标准，计算精确的准确度(记为P@N)：

其中代表由模型预测的前N个临床观察项中，出现在真实的就诊记录中的数目，本实验依次计算了N=5，10和50时的结果来验证提出模型的准确性。

### 结果与分析

在上述实验中，输入每个患者样本d的前次就诊记录来拟合模型，并使用最后一次（第次）入院中实际发生的临床观察项作为评估的依据。此外，还将本模型与最基本的朴素贝叶斯模型（Naive Bayes Model, NBM）进行了对比，结果如表3.3所示。

结果表明，在N取不同值的情况下，本论文提出的模型评估的准确性都高于基准模型，进一步证明了BHMM的有效性。此外，基于这一特点，还可以提出BHMM-DP的进一步应用——临床表现预测，即预测患者治疗轨迹中即将出现的临床观察项。具体来说，给定患者目前的诊疗数据，可以通过公式（3.22）对所有临床观察项出现的概率进行估计，然后根据概率大小进行排序，取前N个作为最有代表性的临床观察项，可以作为患者下次入院身体状况的参考，从而有助于做好早期预防。

表3.3 与NBM比较评估结果的准确性

|  |  |  |  |
| --- | --- | --- | --- |
| 模型类别 | N | | |
| 5 | 10 | 50 |
| BHMM | 0.96 | 0.94 | 0.82 |
| NBM | 0.87 | 0.85 | 0.73 |

## 本章小结

本章提出了一种基于贝叶斯隐马尔可夫模型的疾病演变过程建模方法，对患者的疾病演变过程进行深入的分析，该模型是一种概率生成模型，将患者的疾病状态视为模型中的隐藏变量，将包含患者诊疗信息的EHR数据作为观测序列，由此将患者的疾病演变过程视为一个隐马尔可夫过程，通过解决马尔可夫模型的学习问题估计出患者在疾病演变过程中对应的一系列疾病状态及状态之间相互转移的路线。

本章利用第二章中从中国某三甲医院心内科收集的心衰患者的EHR数据进行了一系列实验，实验结果表明本模型发现的疾病状态具有良好的临床意义和可解释性。

最后，为了对模型的有效性进行验证，本章还通过比较提出的模型和基本的朴素贝叶斯模型在预测临床表现上的准确性，结果表明提出模型的准确性高于基准模型，证明了提出模型的准确性和有效性。

# 展示系统设计与实现

## 整体框架

画一下前端到后端数据库的流程图

## 功能设计

## 系统实现

# 总结与展望

2017年死亡人数来自网页BBC news:

https://www.bbc.com/zhongwen/simp/world-47474256

# 参考文献

1. World Health Statistics 2020: Monitoring health for the SDGs. Retrieved May 18, 2020
2. Kaptoge S , Pennells L , Bacquer D D , et al. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions[J]. The Lancet Global Health, 2019, 7(10).<https://www.who.int/zh/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>
3. 中国心血管健康与疾病报告编写组. 中国心血管健康与疾病报告 2019 概要[J]. 中国循环杂志 2020年9月第35卷第9期.
4. D. Mould. Models for disease progression: new approaches and uses. Clinical Pharmacology & TEHRapeutics, 92(1):125-131, 2012.
5. P.B. Jensen, L.J. Jensen, S. Brunak. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet, 13 (6):395–405, 2012.
6. Richard K. Burt, et al. Effect of Nonmyeloablative Hematopoietic Stem Cell Transplantation vs Continued Disease-Modifying TEHRapy on Disease Progression in Patients with Relapsing-Remitting Multiple Sclerosis: A Randomized Clinical Trial. JAMA, 321(2):165-174, 2019
7. L. Chen, et al. A novel genotype-based clinicopathology classification of arrhythmogenic cardiomyopathy provides novel insights into disease progression. European Heart Journal, doi.10.1093/eurheartj/ehz172, 2019
8. Ahnul Ha, et al. Association of Angle Width With Progression of Normal-Tension Glaucoma, JAMA Ophthalmology, 137(1):13-20, 2019
9. Eduardo De Pablo-Fernandez, Raph Goldacre, Julia Pakpoor, Alastair J. Noyce, Thomas T. Warner. Association between diabetes and subsequent Parkinson disease, Neurology, 91(2):e139-e142, 2018.
10. K. Ito, S. Ahadieh, B. Corrigan, J. French, T. Fullerton, and T. Tensfeldt. Disease progression meta-analysis model in alzheimer's disease. Alzheimer's & Dementia, 6(1):39-53, 2010
11. 沈洪兵、齐秀英主编．流行病学（第8版）．北京：人民卫生出版社，2013.3：54.
12. 胡丹青. 基于电子病历的急性冠脉综合征患者主要不良心血管事件预测[D]. 2017.
13. Yan A T , Jong P , Yan R T , et al. Clinical trial--derived risk model may not generalize to real-world patients with acute coronary syndrome.[J]. American Heart Journal, 2004, 148(6):1020-1027.
14. Jonnagaddala J, Liaw S, Ran P, et at. Coronary artery disease risk assessment from unstructured electronic health records using text mining[J]. Journal of Biomedical Informatics, 2015, 58:S203-S210.
15. Z. Huang, W. Dong, P. Bath, L. Ji, and H. Duan. On mining latent treatment patterns from electronic medical records. Data Mining and Knowledge Discovery, 2015, 29(4):914-949
16. Panigrahi S , Kundu A , Sural S , et al. Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning[J]. Information Fusion, 2009, 10(4):354-363.
17. Wang J , Hu Y , Xiao F , et al. A novel method to use fuzzy soft sets in decision making based on ambiguity measure and Dempster-Shafer theory of evidence: An application in medical diagnosis[J]. Artificial Intelligence in Medicine, 2016, 69(may):1-11.
18. Dutta, Palash. Uncertainty Modeling in Risk Assessment Based on Dempster–Shafer Theory of Evidence with Generalized Fuzzy Focal Elements[J]. Fuzzy Information & Engineering, 2015, 7( 1):15-30.
19. M. Cohen, A. Grossman, D. Morabito, M. M. Knudson, A. Butte, and G. Manley. Identification of complex metabolic states in critically injured patients using bioinformatic cluster analysis. Critical Care, 14(1):R10, 2010.
20. J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via fused sparse group lasso. In KDD, pages 1095-1103. ACM, 2012.
21. J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via multi-task learning. NeuroImage, 78(0):233-248, 2013
22. X. Wang, D. Sontag and F. Wang. Unsupervised learning of disease progression models, KDD’14, August 24-27, 2014, New York, NY, USA
23. Z. Sun, S. Ghosh, Y. Li, Y. Cheng, A. Mohan, C. Sampaio, and J. Hu. A probabilistic disease progression modeling approach and its application to integrated Huntington’s disease observational data, JAIMA Open, 2(1):123-130, 2019
24. C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto. Multistate markov models for disease progression with classification error. Journal of the Royal Statistical Society: Series D (The Statistician), 52(2):193-209, 2003.
25. R. Sukkar, E. Katz, Y. Zhang, D. Raunig, and B. T. Wyman. Disease progression modeling using hidden markov models. In EMBC, pages 2845-2848. IEEE, 2012
26. J.Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann PublisEHRs, San Mateo, Calif., 1988
27. Baum L E , Petrie T . Statistical Inference for Probabilistic Functions of Finite State Markov Chains[J]. Annals of Mathematical Stats, 1966, 37(6):1554-1563.
28. Baum L E , Eagon J A . An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology[J]. Bull.amer.math.stat, 1967, 37(3):360-363.
29. Baum L E , Petrie T , Soules G , et al. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains[J]. Annals of Mathematical Stats, 1970, 41(1):164-171.
30. Rabiner L R . A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proc IEEE, 1989, 77.
31. Teh Y W , Jordan M I , Beal M J , et al. Hierarchical Dirichlet Processes[J]. Journal of the American Statal Association, 2006, 101(December):1566-1581.